STATE OF DATA SCIENCE 2021

ON THE PATH TO IMPACT

ANACONDA.

EXECUTIVE SUMMARY

We conducted the 2021 State of Data Science survey focusing on how data science as a field is growing, the overall trends in adoption from both commercial environments and academic institutions, and what students can do to prepare for the future.

Given that 2020 and 2021 were affected by the COVID-19 pandemic, we took this opportunity to ask questions around how the pandemic impacted work and how organizations invested in the field.

This report dives into the day-to-day of a data professional, the tools and languages they use to be successful, how open source is adopted, the future of work and business impact, and big questions around automation, bias, explainability, and interpretability.

TABLE OF CONTENTS

METHODOLOGY

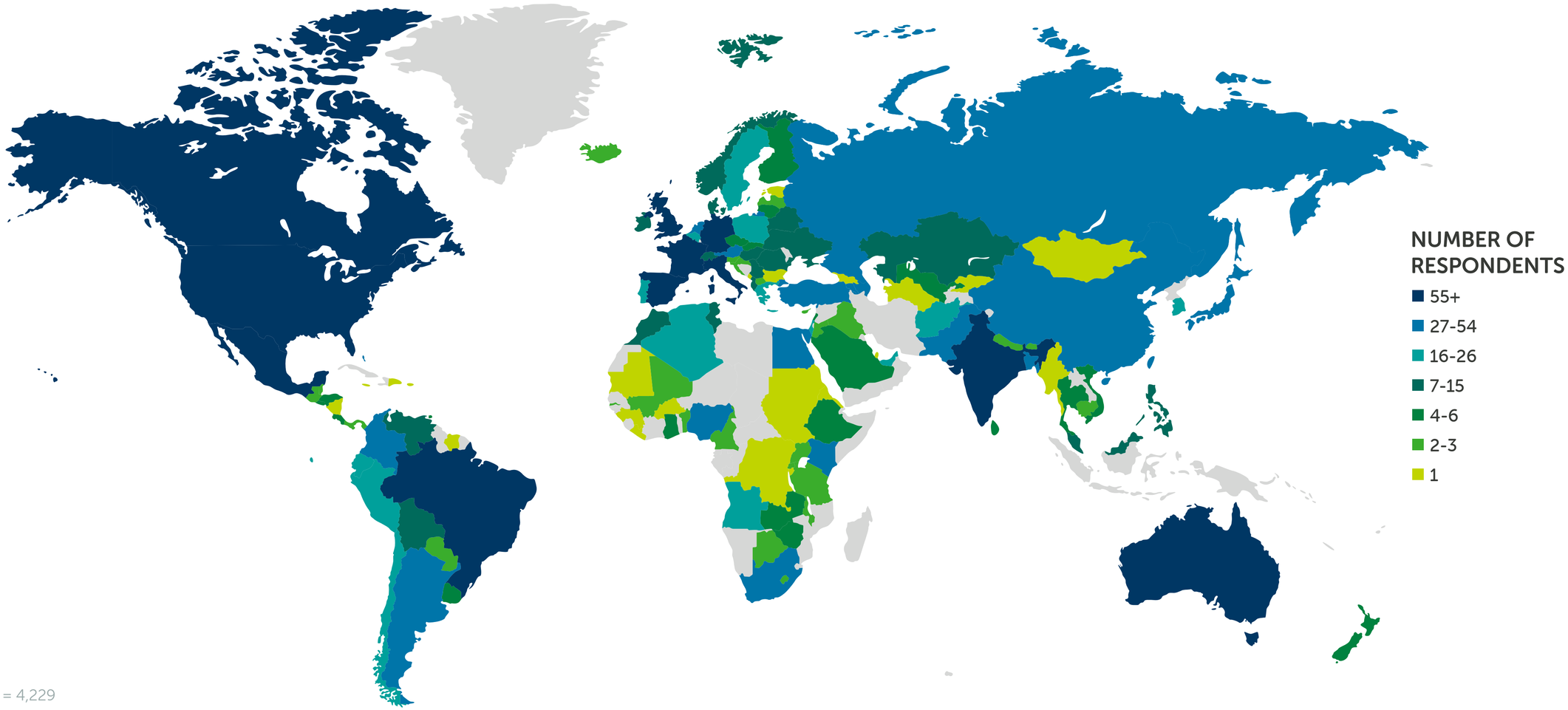
4,299 individuals from more than 140 countries took part in our online survey conducted from April 14, 2021 - May 5, 2021. Respondents came from social media, the Anaconda email database, and Anaconda.org. They had the opportunity to participate in a sweepstakes drawing as an incentive for completing the survey, and two winners were selected at random after the survey was complete. The respondents were divided into three separate tracks: students, academics, and those working in commercial environments. Each of these different cohorts was asked some universal questions, while some questions were unique to each cohort's experience. In the report, we indicate if responses came from either the entire set of respondents or a subset.

Note: All percentages are rounded to the nearest whole percent. Due to rounding, some numbers may not equal 100.

THE FACE OF DATA SCIENCE

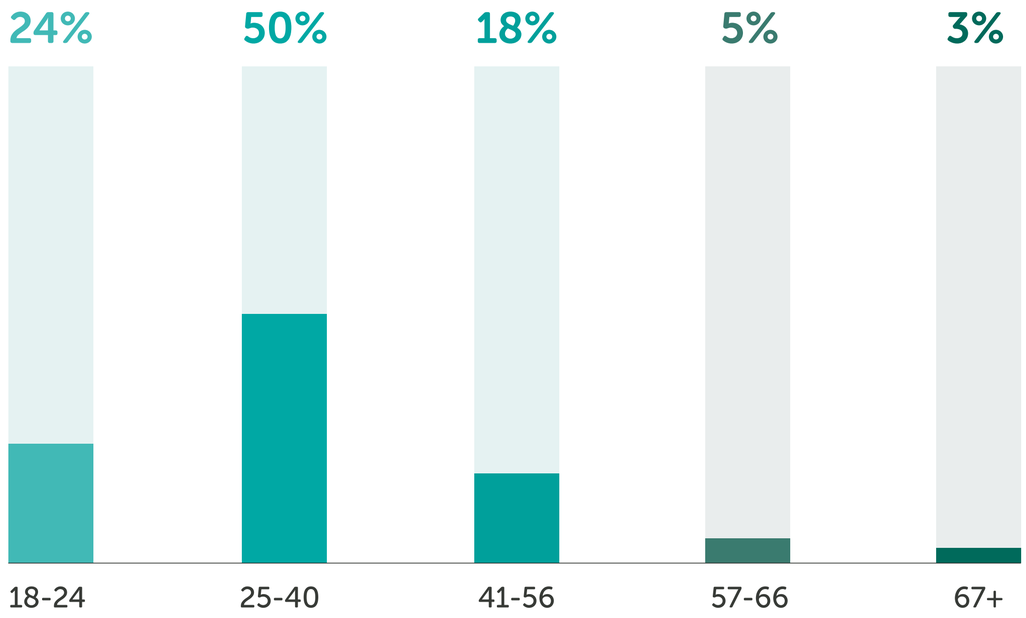
We started this year's survey with a series of questions designed to provide an overview of data professionals. From geographical regions to generational differences, our respondents' demographics give a snapshot of how the data science community is evolving, from job functions, organization size, level of education, and more.

More than 4,200 individuals from 140+ countries participated in our online survey.



n = 4,229

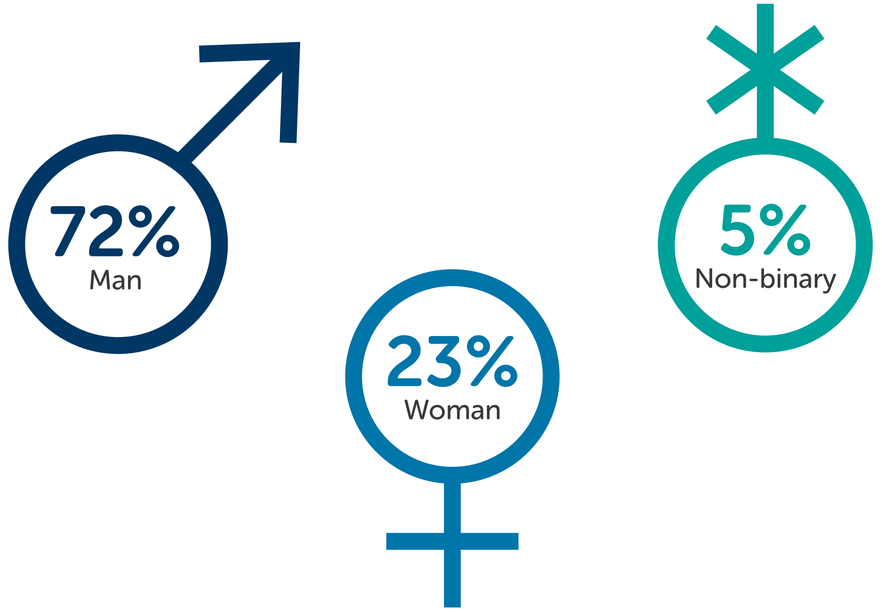
Respondent Age



n = 4,229

Our respondent set skews heavily toward younger generations. Of the 4,229 respondents, 74% are Generation Z (24%) or Millennials (50%). Compared to 2020, this year, we saw a 15% increase in respondents from Generation Z (ages 18-24).

Respondent Gender and How They Identify

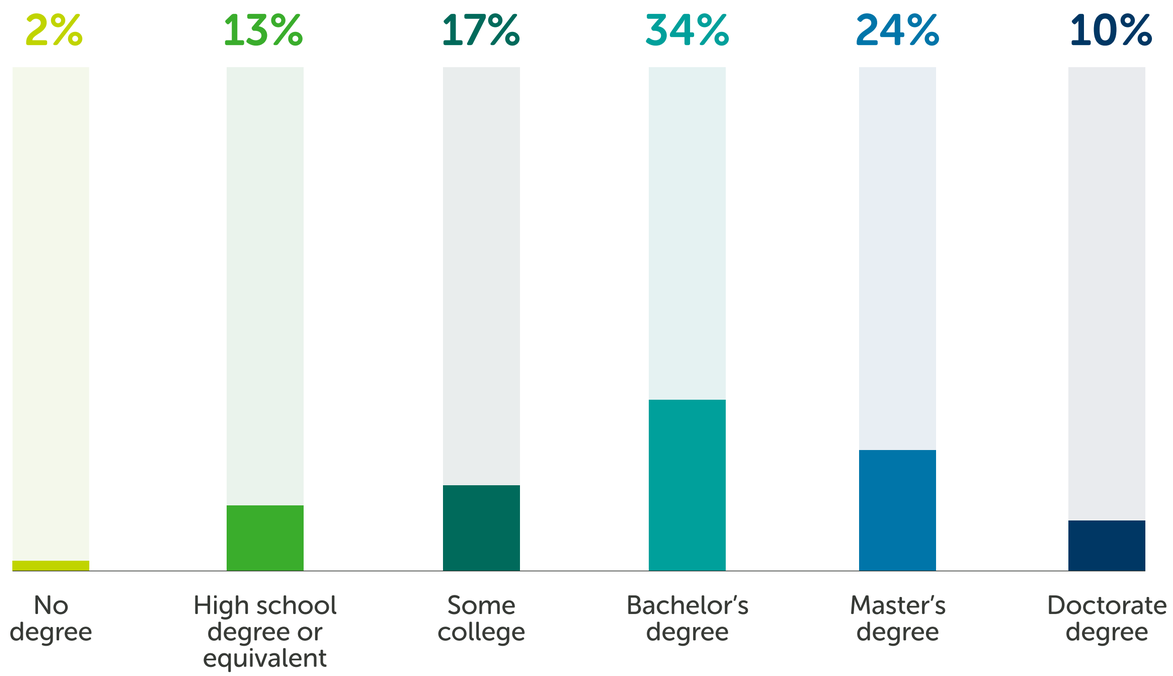


n = 4,229

It comes as no surprise that this data supports what we've seen across other male-dominated STEM-related fields. Nevertheless, there is an opportunity for the industry to work on increasing gender diversity.

Respondent Education Level

The majority of respondents are also well-educated. 68% have a college-level degree.

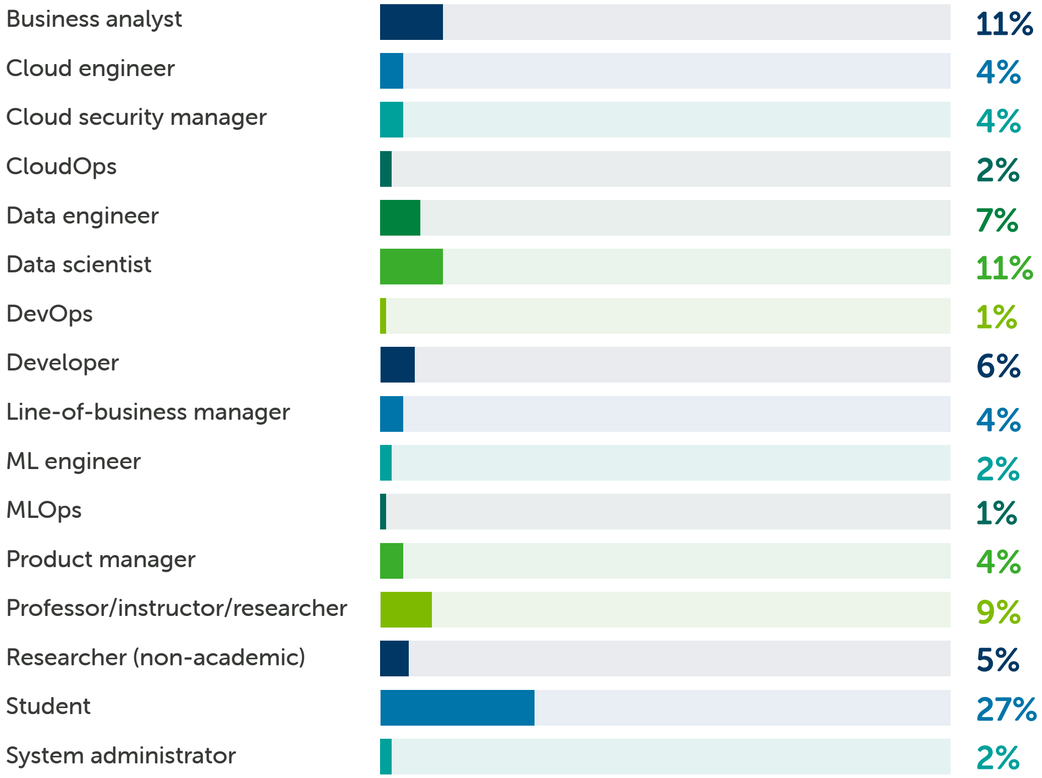


n = 4,229

32%

of respondents do not hold any college degree — with the increasing availability of online courses and ways to build technical skills and experience, having a degree isn't a prerequisite for getting started in data science.

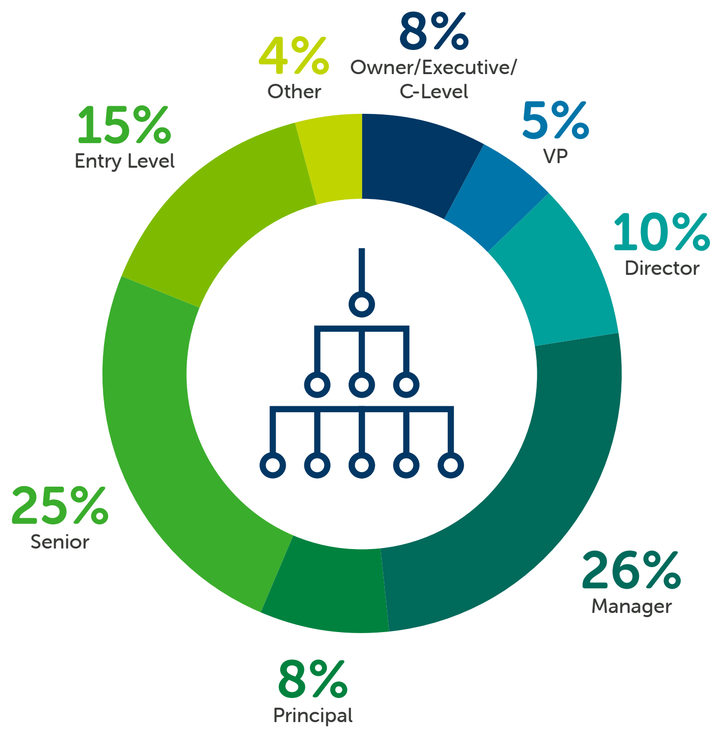
Respondent Primary Job Function



We asked our respondents to choose one role that best represents their job function. However, because there are various data-focused job roles, and on some teams many individual titles are responsible for multiple tasks, there is often overlap between job functions (for example, a Machine Learning Engineer may do product integration, which traditionally may be considered more of an MLOps job).

n = 4,229

Respondent Current Job Level



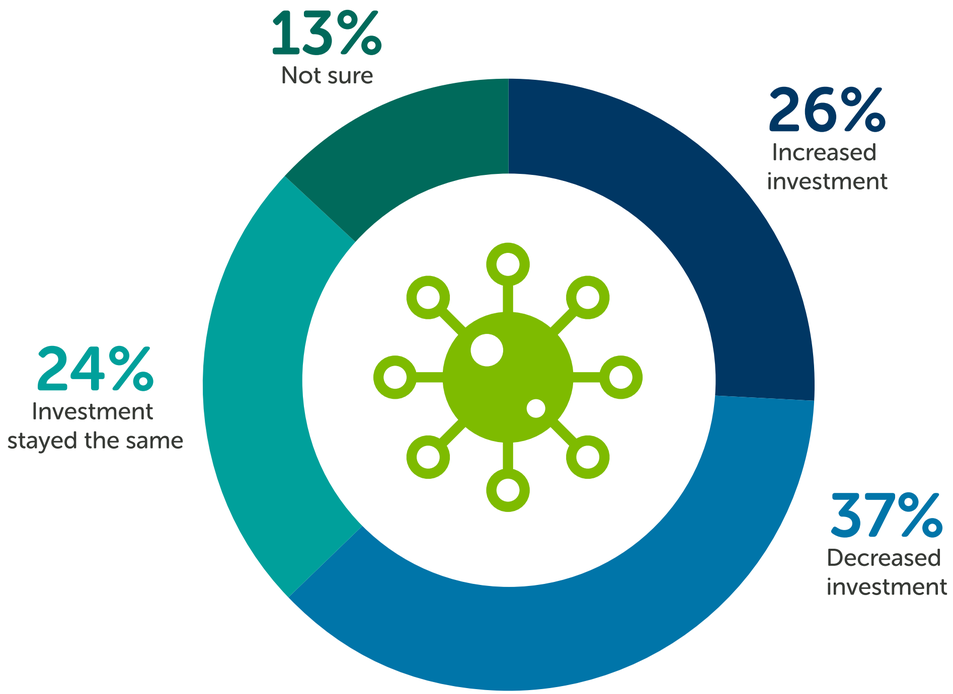
Entry-level positions made up the third-largest group (15%), mirroring the generational demographics of our Millennial and Generation Z respondents who work in a commercial environment.

n = 2,664

HOW HAS COVID-19 AND THE PANDEMIC IMPACTED DATA SCIENCE?

COVID-19 played an important role in how we worked, lived, socialized, and focused our resources over the past year. The pandemic likely influenced answers throughout the survey, especially in allocating resources, job functions, and more.

Did the COVID-19 pandemic impact your organization's investment in data science?



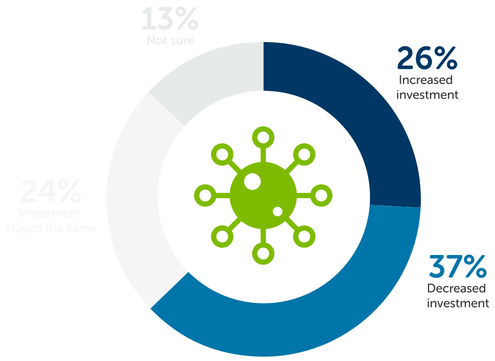
n = 2,229

In total, 50% said their data science investment stayed the same or increased during the pandemic, while 37% saw a decrease.

Of respondents who said the pandemic impacted their organization's investment in data science, 50% said the investment stayed the same or increased, meaning data roles remained significant throughout the pandemic. COVID-19 had a trickle-down effect that impacted virtually every industry – from healthcare to government, financial institutions, and more; they all needed to find ways to act quickly on data and find solutions to new problems. Additionally, when asked how involved their role is in business decisions, 14% of respondents said "all" decisions rely on insights interpreted by them or their team, and 39% said "many" business decisions rely on them. While there is still work needed to ensure we bring data scientists into the fold, it's encouraging to see their value is recognized in organizations and might explain why the field avoided a sharp decrease in investment. We dive into this more in the report's data literacy and business impact section and identify further opportunities for improvement.

We took a deeper look at why individuals said there was an increase or decrease in investment.

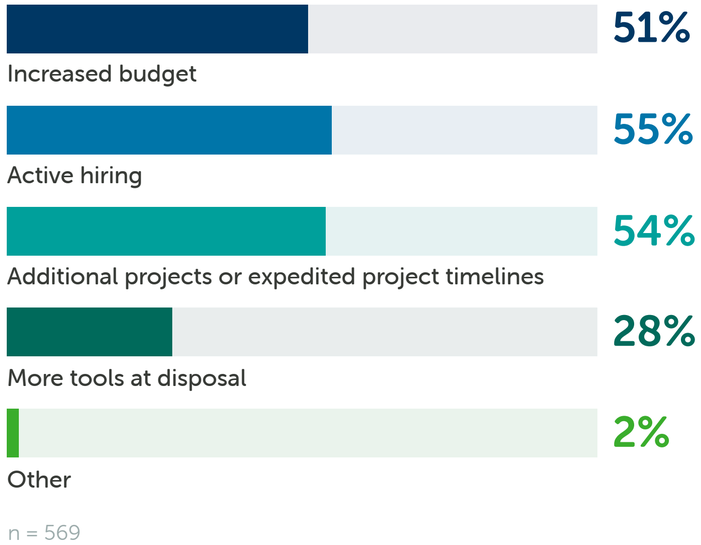
We asked respondents to select all that apply.



Increased Investment

Those who said their investment increased saw it mostly allocated toward hiring, followed by increasing budgets and additional or expedited projects.

In what ways has your organization increased its investment in data science?

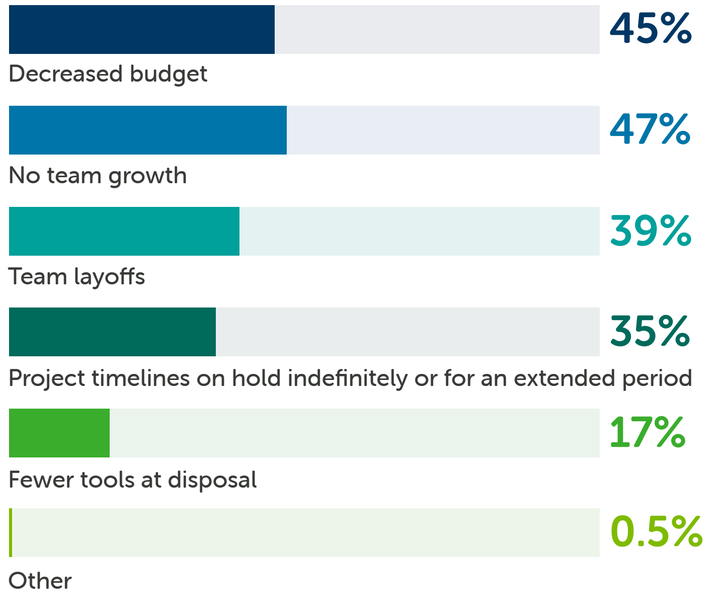


n = 569

Decreased Investment

For those who saw a decrease in investment, the trends were toward no team growth, followed by decreased budgets.

In what ways has your organization decreased its investment in data science?



n = 818

DATA PROFESSIONALS AT WORK

The majority of our respondents work in commercial environments. We took a closer look at these responses to get more detail on where data professionals sit in the organization, how they spend their time, what tools they use, and the most significant challenges in their roles.

Respondent Industry

Technology	10%	Manufacturing	3%
Academic	7%	Aerospace	2%
Consulting	7%	Chemicals	2%
Automotive	6%	Defense	2%
Banking	6%	Energy	2%
Engineering	5%	Retail	2%
Finance	5%	Telecommunications	2%
Apparel	4%	Entertainment	1%
Education	4%	Environmental	1%
Communications	4%	Food & Beverage	1%
Agriculture	3%	Insurance	1%
Biotechnology	3%	Media	1%
Construction	3%	Not For Profit	1%
Electronics	3%	Pharmaceutical	1%
Government	3%	Transportation	1%
Healthcare	3%	Utilities	1%

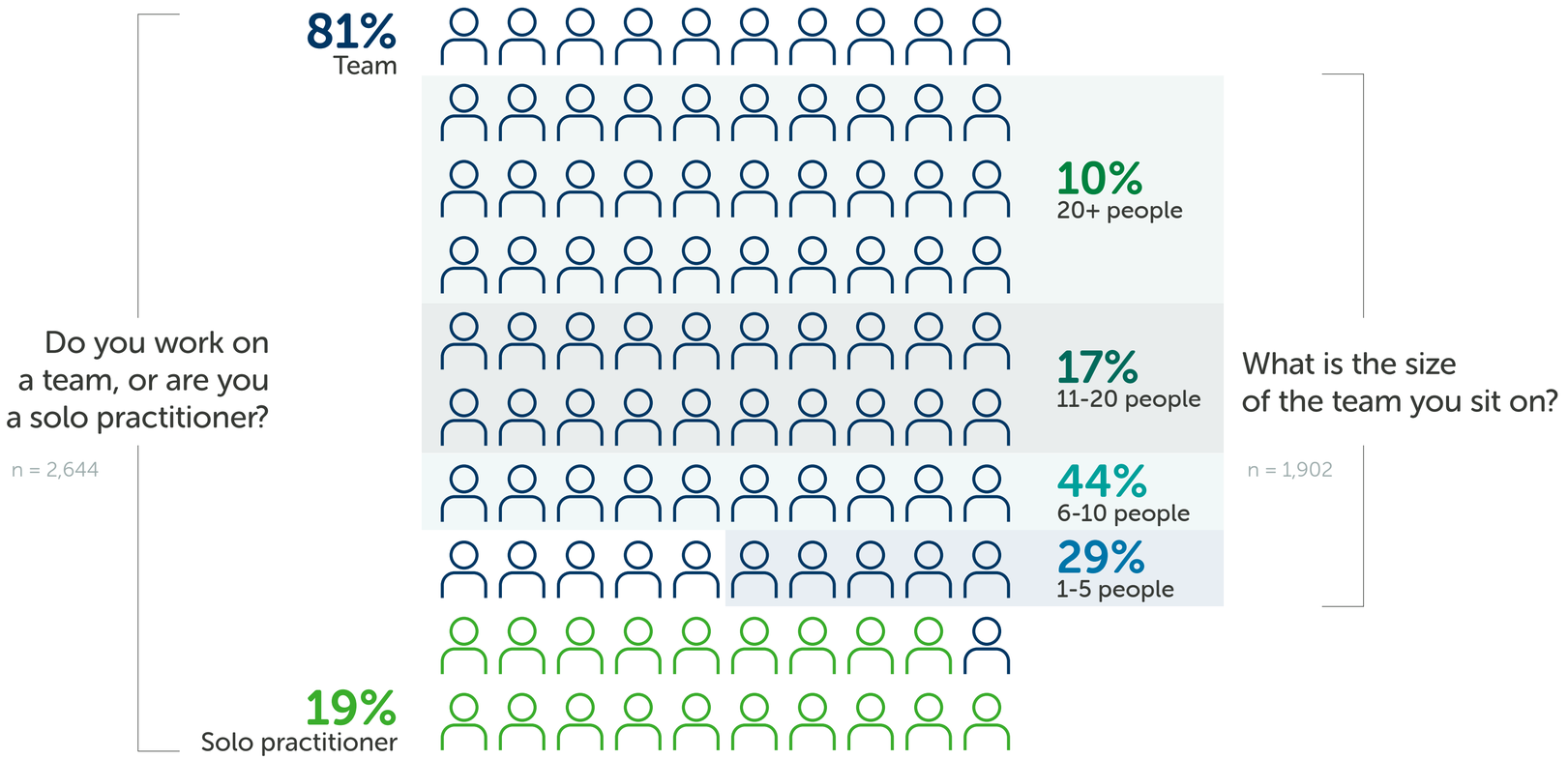
n = 2,644

Of the individuals surveyed, the top nine industries represented were: Technology (10%), Academic (7%), Consulting (7%), Automotive (6%), Banking (6%), Engineering (5%), Finance (5%), Apparel (4%) and Education (4%).

We saw representation from virtually every industry in our commercial respondent set. From higher education to not-for-profit and retail, companies are prioritizing data-driven roles.

Team Size

To better understand data professionals at work and how they fit into the corporate structure, we asked individuals about the size of their organization and if they worked on a team or by themselves.

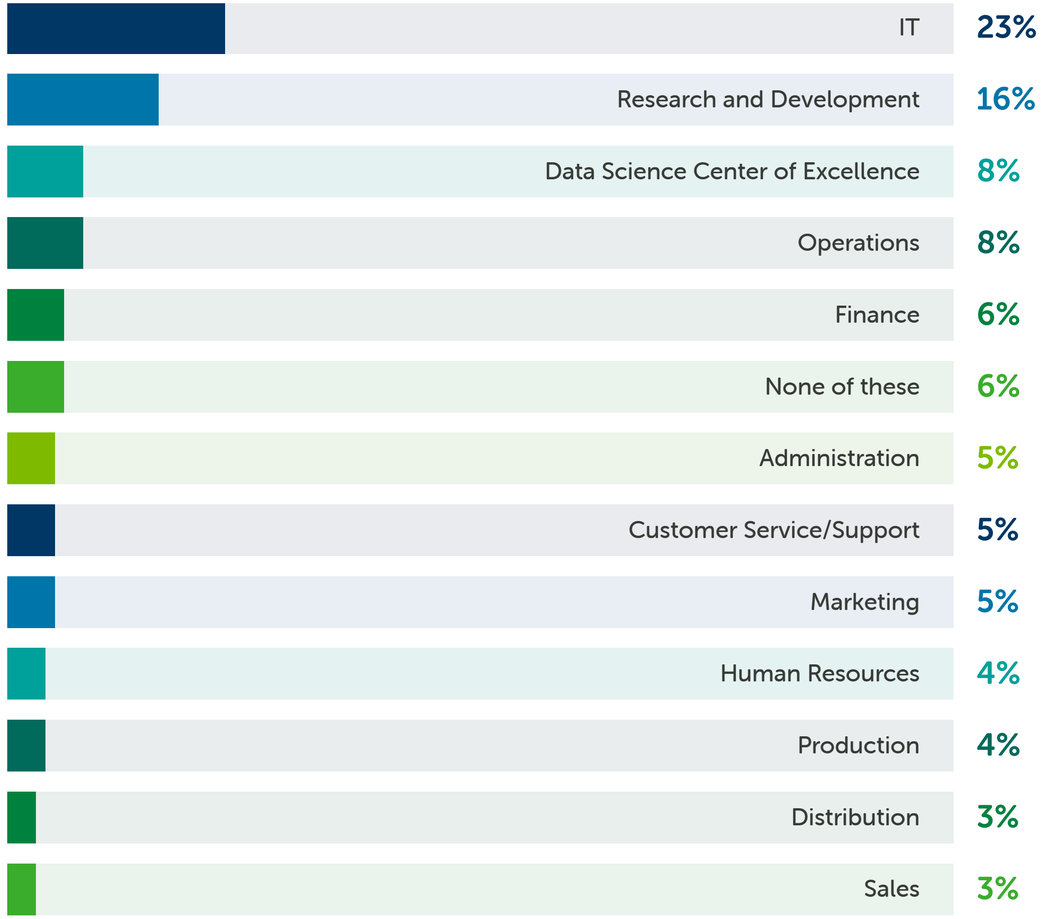


While 81% of our respondents work on a team, most teams tend to be small. For example, 73% of respondents indicated they were part of a team with ten people or fewer. While hiring data science talent is difficult, new challenges will emerge as leaders decide how to structure these teams. The best approach will depend on the stage of maturity of an organization's data science or AI program.

What department does your role fall under?

Where do data roles fall within the organization? The short answer is: everywhere. Organizations structure their departments differently. Sometimes there is an entire data-focused team; other times, data scientists are seated within different departments. Most of our respondents' roles fall under IT (23%) and Research & Development (16%). In large organizations with 10,001+ employees, we found that individuals in data-focused roles primarily fall under R&D (18%), IT (18%), and finance departments (16%). In larger organizations with more mature data science programs, there are more resources to ensure data roles are integrated throughout the business and in cross-functional teams.

For data scientists specifically, they are most often working under the IT function of the organization (22%), followed by R&D (21%) and the Data Science Center of Excellence (20%). We found that 43% of organizations with a Data Science Center of Excellence have data science teams made up of 1-5 people, compared to IT teams with data scientists, which most of our respondents said have 6-10 people. This is likely because IT teams also include additional roles such as system analysts, developers, and programmers.



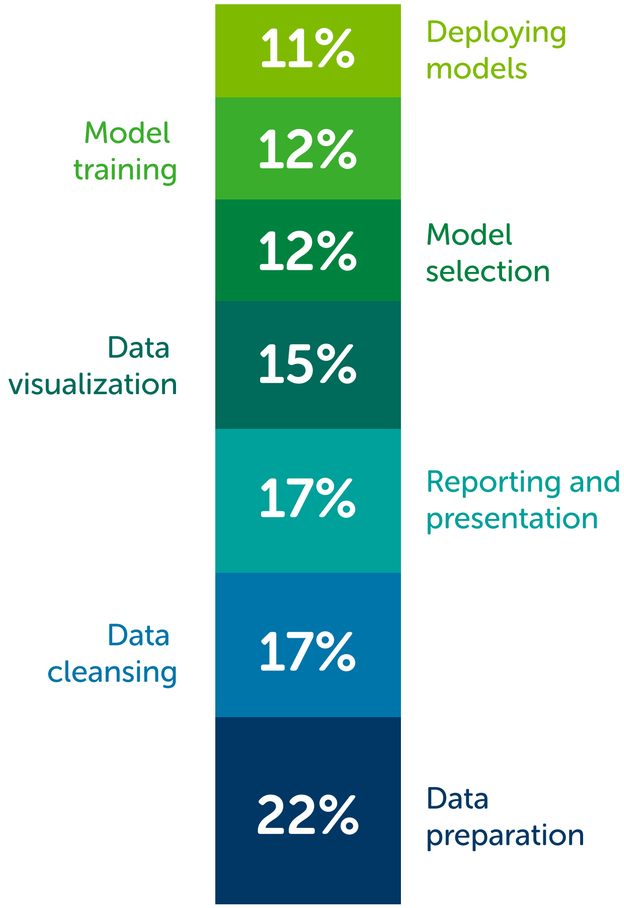
n = 2,644

How do data scientists spend their time?

Data scientists spend their day focused on various tasks that require a diverse set of technical and non-technical skills. When asked how much time they spend on tasks, respondents stated they spent about 39% of their time on data prep and data cleansing, which is more than the time spent on model training, model selection, and deploying models combined.

While data preparation and data cleansing are time consuming and potentially tedious, automation is not the solution. Instead, having a human in the mix ensures data quality, more accurate results, and provides context for the data.

Beyond preparing and cleaning data, interpreting results is also critical. Visualization and demonstrating the data's value through reporting and presentation are essential parts of making that data actionable and ultimately providing answers to critical questions.



n = 2,030

We asked our respondents how much time they spend on each of the above tasks, and for each item, enter a number representing the percentage of time spent on each task relative to the other tasks on this list. This is the average of the reported percentages.

Getting to Production

Data can provide immense value to an enterprise, which goes beyond gathering business insights, refreshing dashboards, and monitoring KPIs. Organizations that deploy machine learning models and other data science outputs to power business functions and products have a competitive advantage and rely on the value data scientists bring to the table. However, while getting models to production is one of the most rewarding jobs of a data professional, these individuals are often faced with challenges outside of their control.

From the survey responses, 28% said they do not deploy models in production. However, the respondents in commercial environments who do move models to a production environment cited these top four roadblocks:

27% Meeting IT security standards

24% Re-coding models from Python/R to another language

23% Managing dependencies and environments

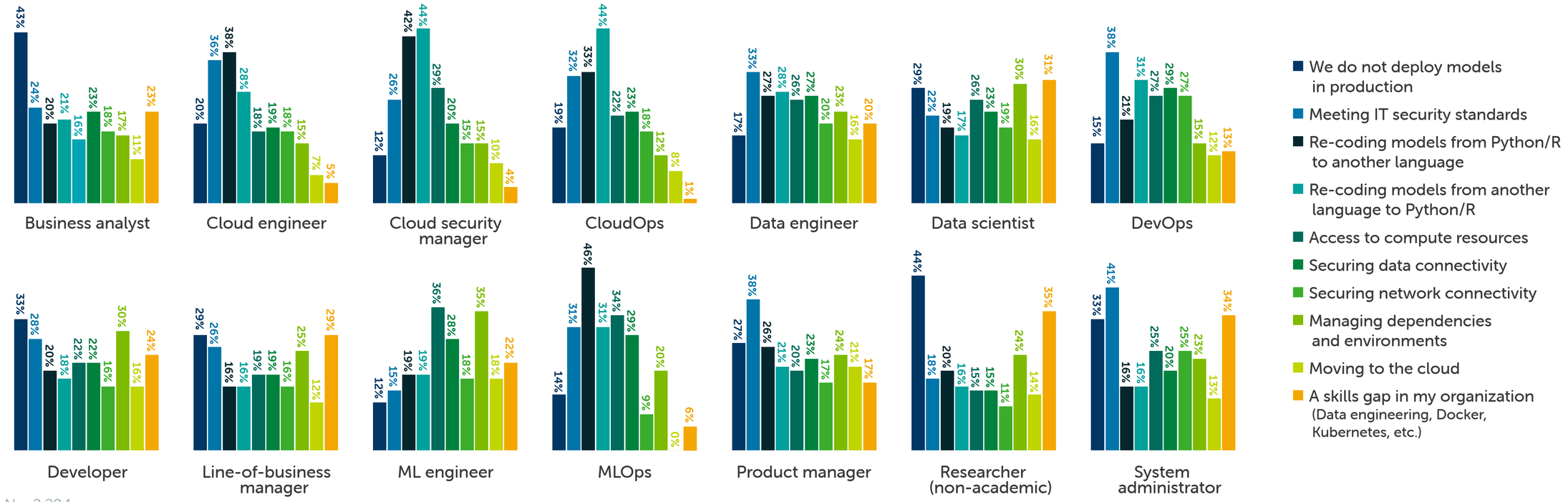
23% Re-coding models from another language to Python/R

Barriers to production differ across job roles and departments. Alignment on a project's purpose, scope, budget, and goals across the organization and amongst technical and business stakeholders is critical to getting models deployed. Organizations have an opportunity to address these roadblocks to see an increased impact from their data science and machine learning efforts.

What roadblocks do you face when moving your models to a production environment?

We asked respondents to select all that apply.

When we break the roadblocks down by job function, there is a clear difference between the priorities based on role in the business. For example, data scientists see their biggest roadblock as a skills gap, which makes sense given the gaps we see between what enterprises need and what educational institutions teach.



N = 2,294

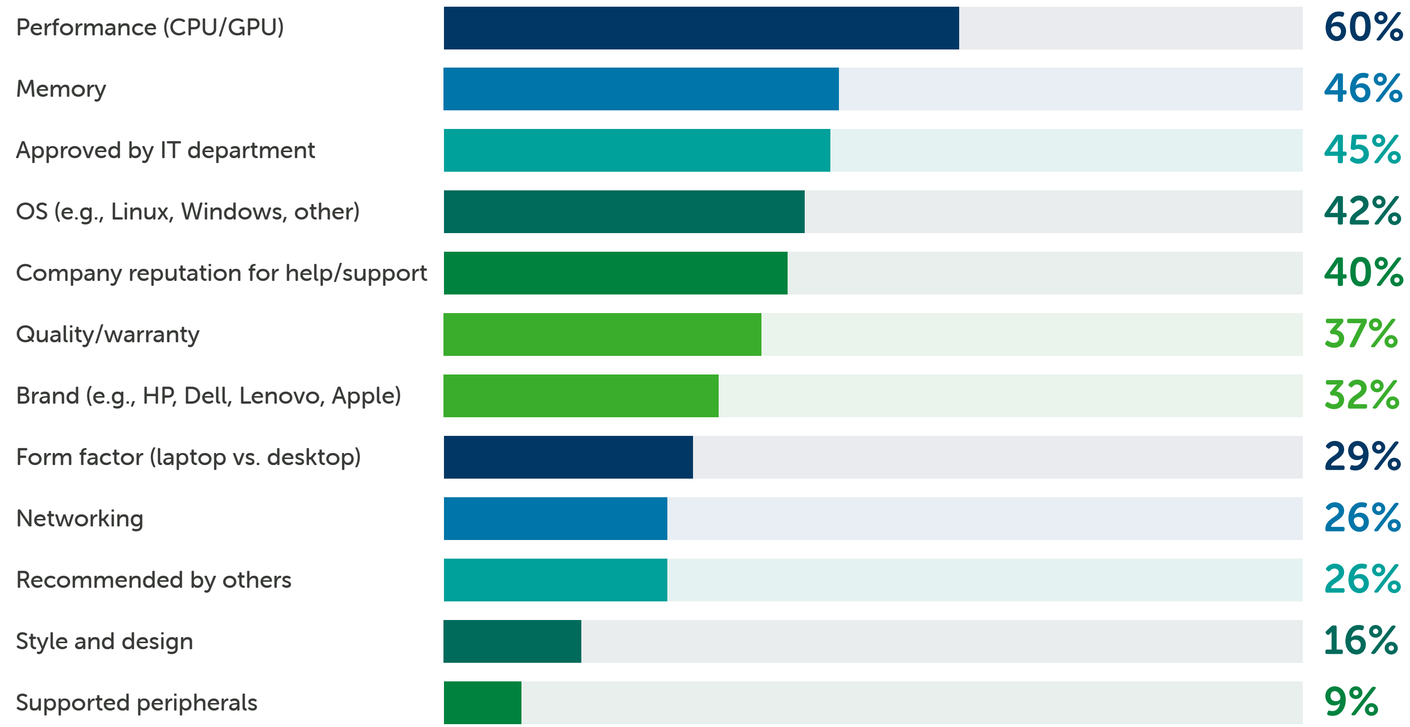
Meeting IT security standards is the top blocker for DevOps, Data Engineers, Product Managers, and System Admins.

Re-coding models was the biggest roadblock to production for those involved in infrastructure.

For Machine Learning Engineers, the biggest roadblock to getting models to production is access to compute resources.

When purchasing a system for your data science workflows, which of these factors most influence your decision?

We asked respondents to choose their top four factors.



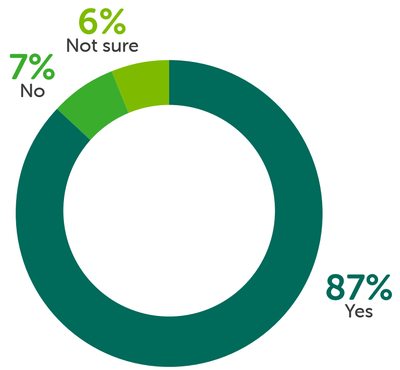
n = 3,104

Additionally, when asked about purchasing a system for data science workflows, 60% said performance (CPU/GPU) is a top factor when deciding between hardware options. Machine Learning Engineers view compute resources as the most significant roadblock to deploying models to production. This is something enterprises should be mindful of when choosing hardware and cloud resources.

ENTERPRISE ADOPTION OF OPEN SOURCE

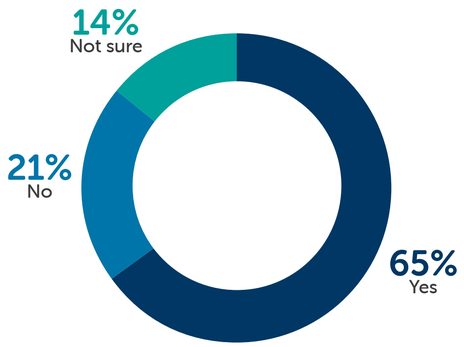
Using and contributing to open-source software (Python / R libraries such as pandas, NumPy, etc.) is a key differentiator of the most innovative organizations. By using open-source software, organizations can save tremendous amounts of time and resources. Rather than purchasing from a single vendor or building everything in-house, this crowdsourced model leverages multiple minds at work to accelerate projects that would typically take years.

Does your organization allow the use of open-source software?



n = 2,289

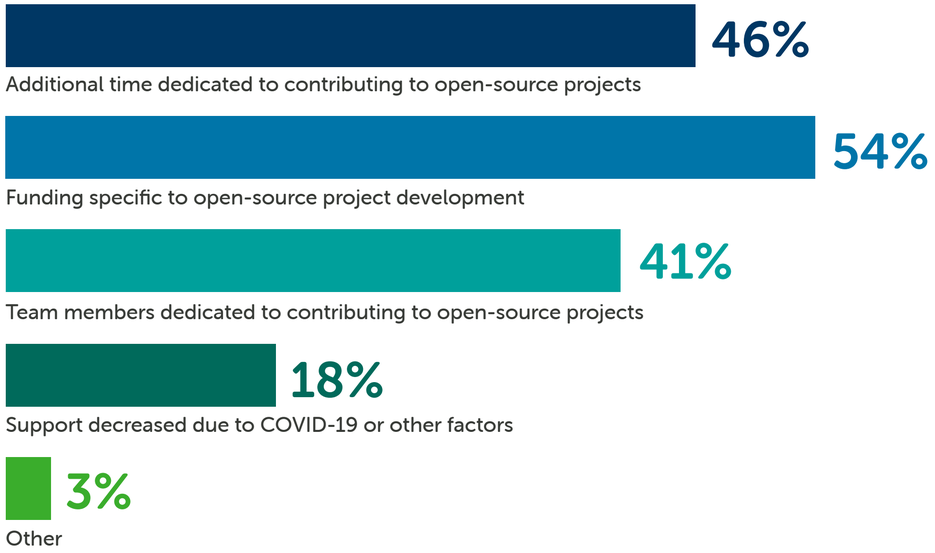
Does your employer encourage you and your team to contribute to open-source projects?



n = 2,107

How does your employer empower you and your team to contribute to open source?

We asked respondents to select all that apply.



n = 1,366

Of the 65% of individuals who said their teams were encouraged to contribute to open-source projects, the majority of respondents (54%) said that their employers are empowering them to contribute to open source through an increase in funding related to open-source project development. This was followed closely by respondents saying they receive additional time dedicated to contributing (46%), and then team members dedicated to contributing to open-source projects (41%). Even as the past year has presented its share of challenges, only 18% of survey respondents said that employer support for open source decreased due to COVID-19 or other factors.

Securing the Open-Source Pipeline

How organizations ensure open-source packages used for data science and machine learning are secure and meet enterprise security standards.

We asked respondents to select all that apply.

n = 1,972

When asked about how those organizations that use open-source software ensure security, 45% of respondents say they use a managed repository, 30% use a vulnerability scanner, and 33% do manual checks against a vulnerability database. However, 25% are not securing their open-source pipeline, and 20% did not report any knowledge about open-source package security.

Like any proprietary software, open source also comes with inherent security management challenges. However, open source's nature of being "open" means various individuals are involved in the process, making it vulnerable to risk. The good thing is that it also allows contributors and maintainers to catch and patch any vulnerabilities quickly. Organizations must understand that they can reap the value and innovation of open-source software while being mindful of pipeline management and security.

Securing the Open-Source Pipeline

For the organizations that aren't using open-source software, what is the barrier to entry? There are various reasons, but similar to the roadblocks for deploying models to a production environment, security concerns reign supreme.

What roadblocks are preventing your organization's use of open source?

We asked respondents to select all that apply.

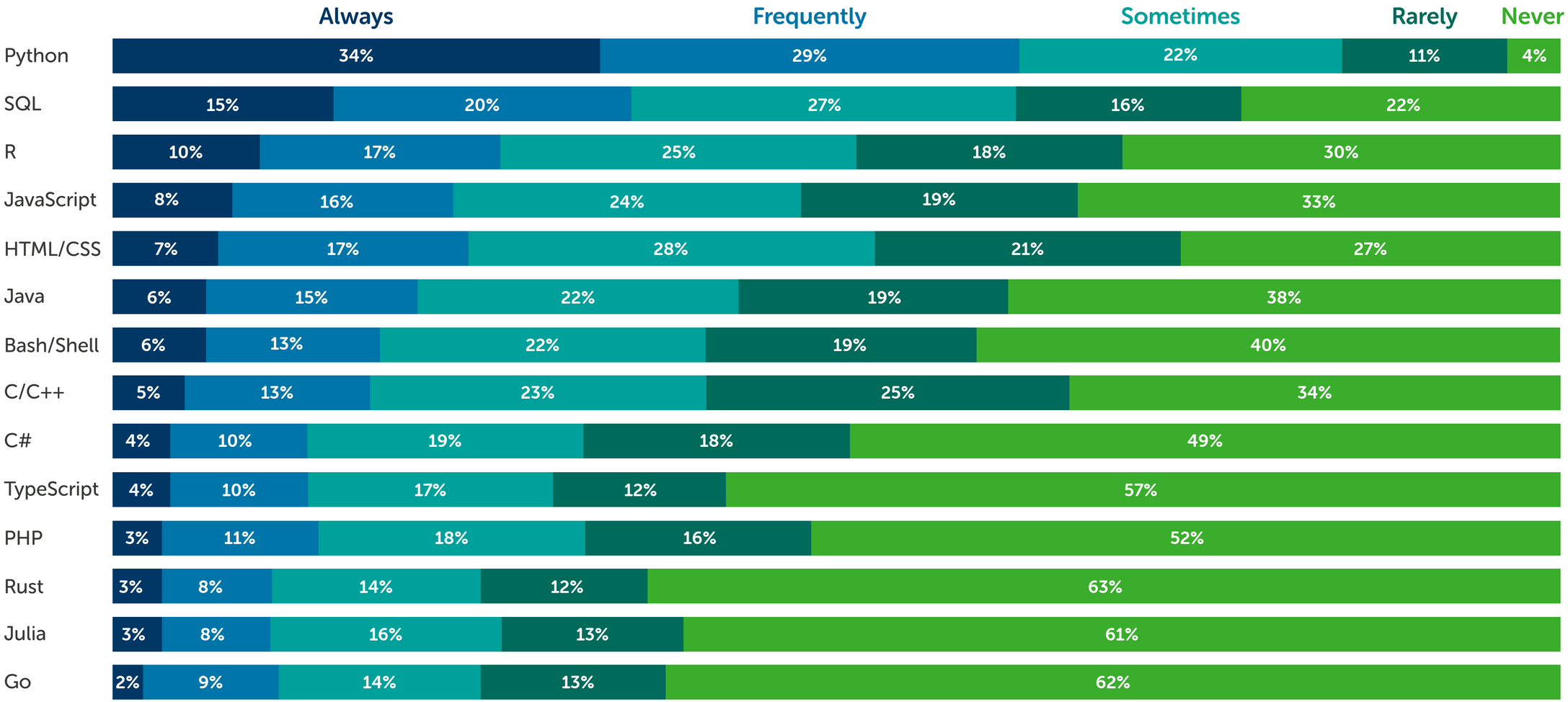
n = 142

Based on the responses, a majority of organizations are using open-source software. But, of the 7% of respondents who are not, the biggest reason why is fear of CVEs (Common Vulnerabilities and Exposures), potential exposures, or risks. These fears also reflect the 25% of respondents who aren't securing their open-source pipeline and 30% of respondents who lack knowledge or understanding of open source. These concerns are particularly prevalent among mid-to-large-sized organizations, perhaps reflecting the more stringent regulatory environment in which many of these companies operate.

POPULARITY OF PYTHON

For data scientists, researchers, students, and professionals worldwide, Python is becoming an increasingly popular programming language.

How often do you use the following languages?



n = 3,104

How often do you use the following languages?

Python appears poised to continue its dominance in the field. 63% of respondents said they always or frequently use Python, making it the most popular language included in this year's survey. In addition, 71% of educators are teaching Python, and 88% of students reported being taught Python in preparation to enter the data science/ML field. Even in our own Anaconda usage data, we've seen impressive growth in Python. Between March 2020 to February 2021, the pandemic economic period, we saw 4.6 billion package downloads, a 48% increase from the previous year. We believe some of this increase could be related to workers transitioning to work from home and more individuals having free time during the pandemic to learn, improve their skills, and pursue their interest in Python.

Beyond being a top language used in commercial environments and taught at universities, Python's popularity can also be demonstrated by various other factors, such as its ease of use, libraries, and community. 20% of students said the biggest obstacle to obtaining the experience required for a career in data science is learning a new language. With most educators teaching Python and Python's continued popularity in the data science community, there is an opportunity for the Python language to become an industry standard. Standardization could help solve re-coding pain points associated with deploying models into production.

When asked which data science and machine learning tools organizations use, it's no surprise given our survey sample; 51% of our commercial respondents said they use Anaconda. Other popular tools include GitHub (35%), Azure ML Studio (23%), Power BI (21%), and Tableau (20%).

DATA LITERACY AND BUSINESS IMPACT

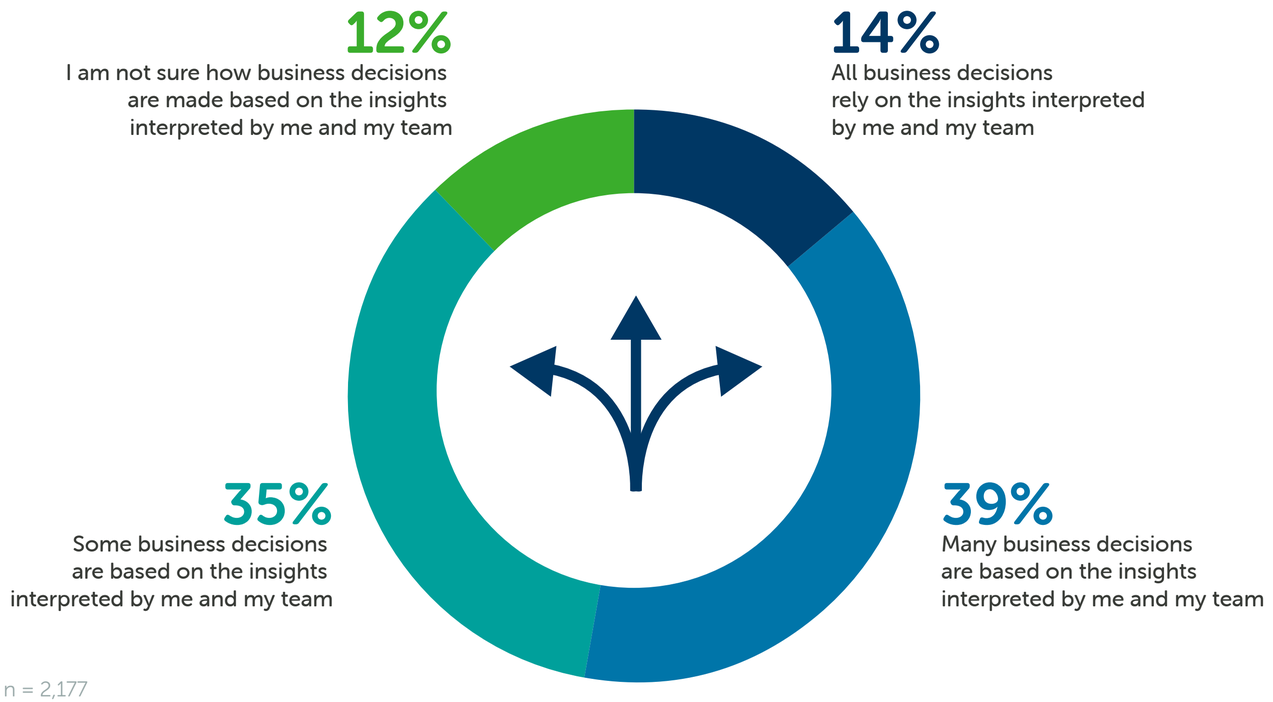
Organizations rely heavily on data to help make decisions, but interpreting and communicating that information is what ensures it's understood and appropriate action happens. From our results, we learned business leaders have an opportunity to become more data literate, and data scientists can improve their business skills to impact how decisions are made.

A lack of business knowledge is impacting a data professional's role in business decisions.

Data practitioners are often brought into business decisions, but there is a gap in how their insights are ultimately used. For example, 39% said many decisions are based on insights interpreted by me and my team. While this shows the growing maturity of the field and might explain why many organizations avoided cuts to these departments during COVID-19, there's still room for growth.

Data science teams bring valuable knowledge to the table, but they don't always have the business background or experience to provide context. In addition, the gap in data literacy amongst leaders makes it difficult for them to enact change based on their insights. There needs to be a greater connection between the C-suite and its data practitioners.

How involved our respondents are in business decisions



n = 2,177

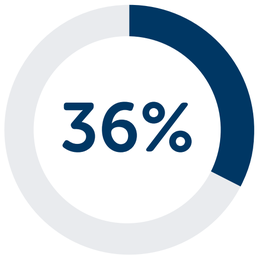
DATA LITERACY AND BUSINESS IMPACT

Data literacy is critical to making impactful data-driven business decisions. Yet, while most leaders have a baseline of data knowledge, there is still a gap. Based on our responses, only 36% of people said their organization's decision-makers are very data literate and understand the stories told by visualizations and models. In comparison, 52% described their organization's decision-makers as mostly data literate but needing some coaching on the stories told by visualizations and models.

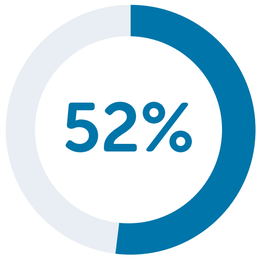
This lack of data literacy at the executive level ultimately hurts the ability to make data-driven business decisions. Additionally, 25% of respondents said that a lack of data literacy among decision-makers at their organization limited their team's ability to impact business decisions.

Other factors limiting data scientists' ability to impact business decisions are that priorities shift quickly (33%), there are not enough resources for effective analysis (25%), and that they or their team cannot effectively demonstrate business impact (11%).

Data literacy of leadership and skills that are missing in the data science/ML area of the organization



My organization's decision-makers are very data literate and understand the stories told by visualizations and models



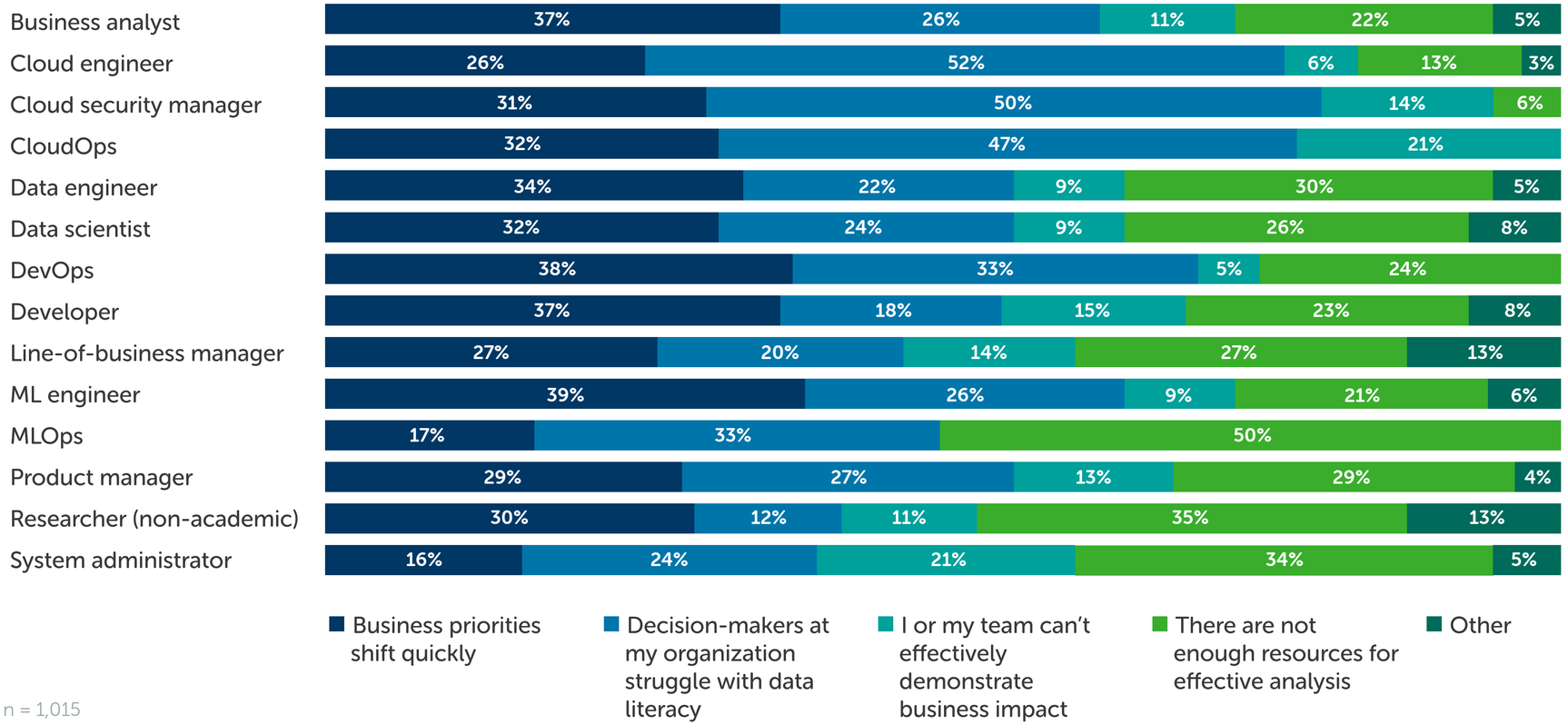
My organization's decision-makers are mostly data literate but need some coaching on the stories told by visualizations and models

My organization's decision-makers have trouble understanding the stories told by data visualizations and models

n = 2,177

What's limiting our respondents' ability to impact business decisions based on their job function?

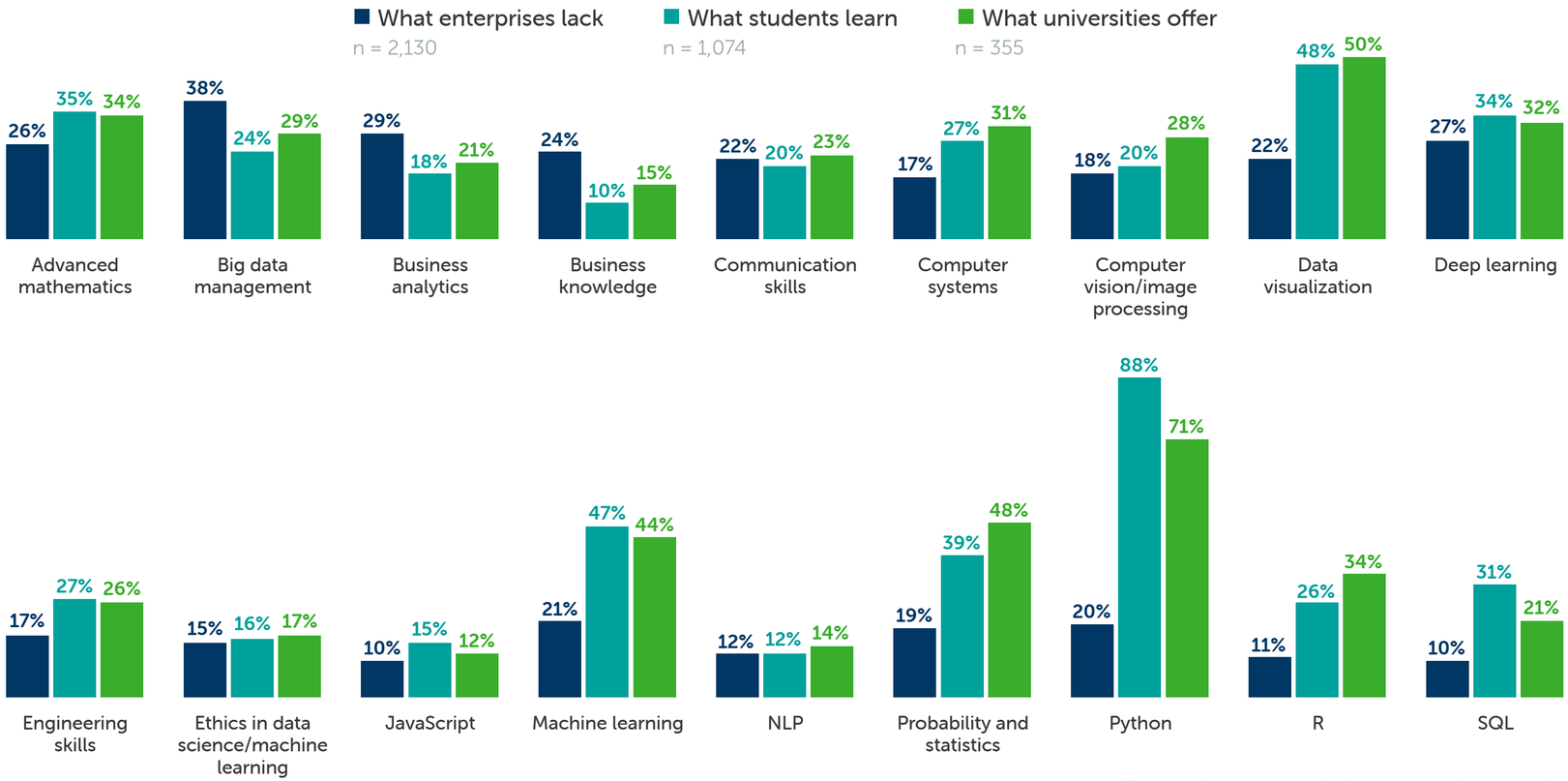
When breaking this question down by job function, there were trends between different roles and what's limiting their ability to impact business decisions. Cloud Engineers, Cloud Security Managers, and CloudOps all felt their most significant limitation was decision-makers struggling with data literacy. Business Analysts, Data Scientists, Machine Learning Engineers, and more felt their limitations stemmed from business priorities shifting quickly. Additionally, MLOps indicated that they do not have enough resources, mirroring their biggest roadblock to deploying models: compute resources.



n = 1,015

Skills gaps and how they're impacting decision making

Our survey shows gaps between what enterprises need, what institutions are teaching, and where students feel the most confidence in their skills. For example, nearly a quarter of enterprise respondents listed "business knowledge" as lacking from data practitioners at their organization. "Business knowledge" also ranked low on questions about what data science students are learning (10%) and being taught in school (15%).



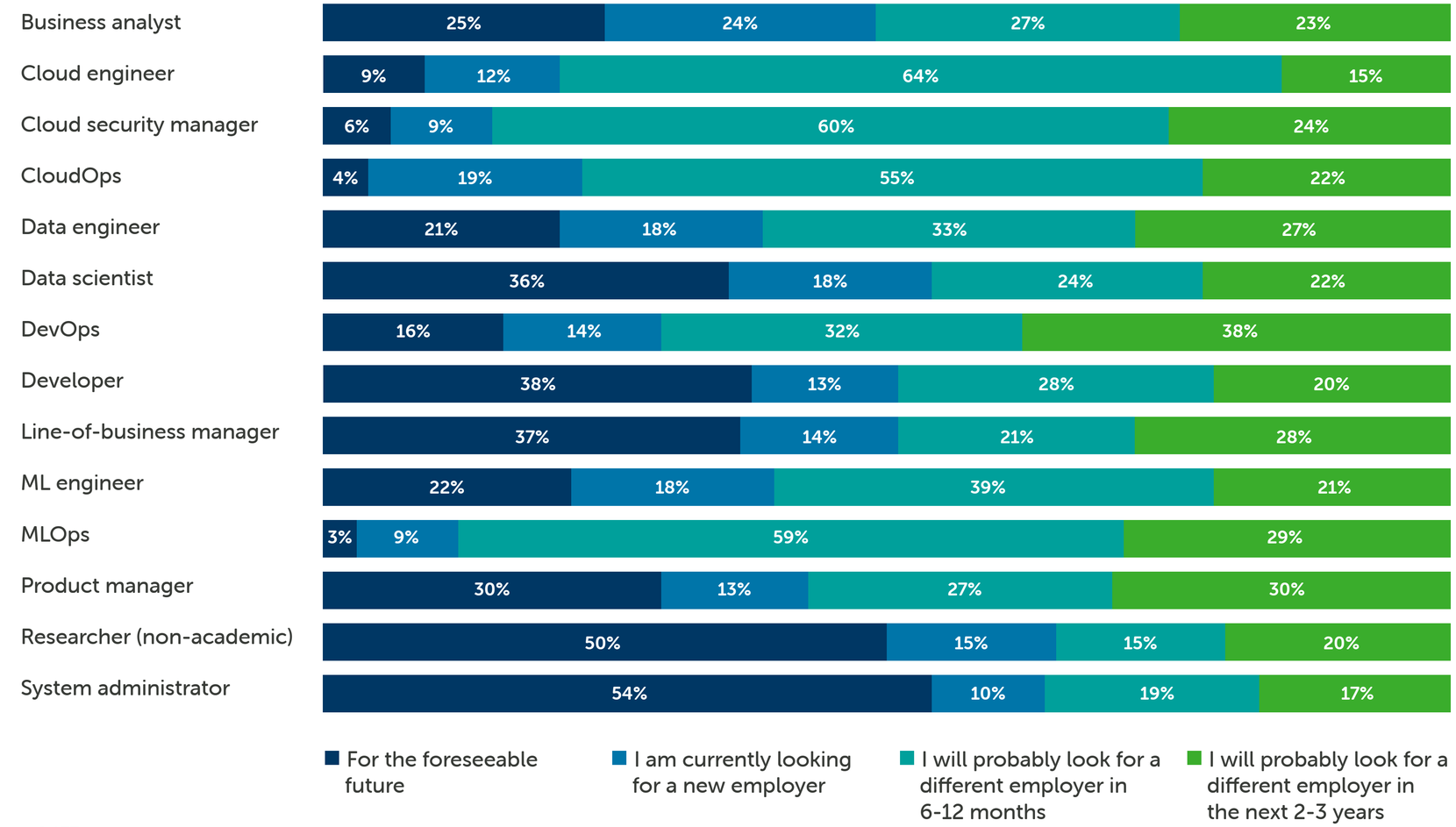
Respondents were asked to select all that apply.

Overall, soft and business-related skills were the most significant gaps between what universities teach and what organizations need. Being involved in strategic business conversations and communicating and explaining results to stakeholders are skill sets that can bridge the gap between data literacy and decision making.

DATA JOBS AND THE FUTURE OF WORK

We found that job satisfaction correlates with job function in the data science field. Pain points that interfere with being able to do jobs effectively are why individuals seek new positions. According to the Bureau of Labor Statistics, data science is listed in the top 15 fastest-growing occupations and is projected to have a 31% job growth over the next 10 years. Considering there are so many organizations looking for data roles coupled with the fact that there is an overall talent shortage for these technical positions, organizations need to ensure they're meeting the needs of their employees to attract and retain talent. Whether from not having enough resources to do their job effectively or a data literacy gap in leadership, there is potential for a high rate of employee churn in the 1-2 year horizon.

How long our respondents plan to stay with their current employer



n = 2,177

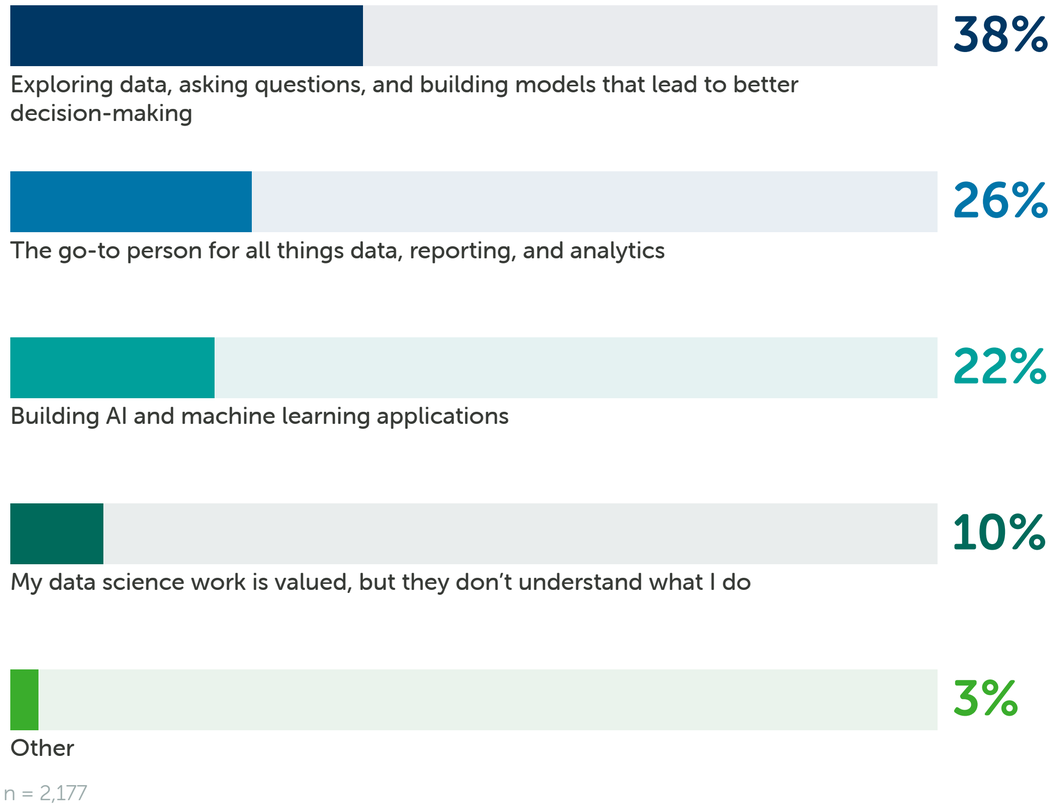
32%

of our total respondents are looking for a job in the next 6-12 months.

Of our respondents, business analysts are most likely to look for a new job right now. In the next 6-12 months, Cloud Engineers, Cloud Security, Cloud Ops, and ML Ops will most likely look for a new job. Data Scientists, Developers, System Admins, Line-of-Business-Managers and Researchers are most likely to stay in their current positions and are the most satisfied with their work. When asked about how leadership perceives our respondents' roles, 10% of respondents said, "My data science work is valued, but [leadership] doesn't understand what I do." Business readiness, having the right resources, as well as perception of data science roles versus the reality could impact overall satisfaction at work.

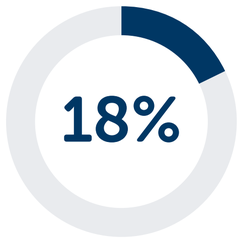
Which of the following do you think best describes how leadership perceives your role?

"Exploring data, asking questions, and building models that lead to better decision-making" was our top response for how leadership perceives data roles. Yet, while data practitioners are involved in decisions, they may not always feel their insights are accurately considered.

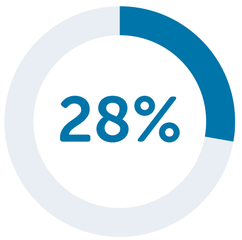


n = 2,177

Our student respondents shared their biggest obstacles to obtaining experience required for a career in data science or related fields.



Cost

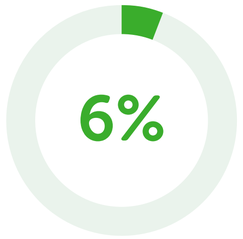


Finding an internship

Learning new languages

The saturated field makes it difficult to stand out

Lack of resources due to COVID-19



Other

n = 1,074

28%

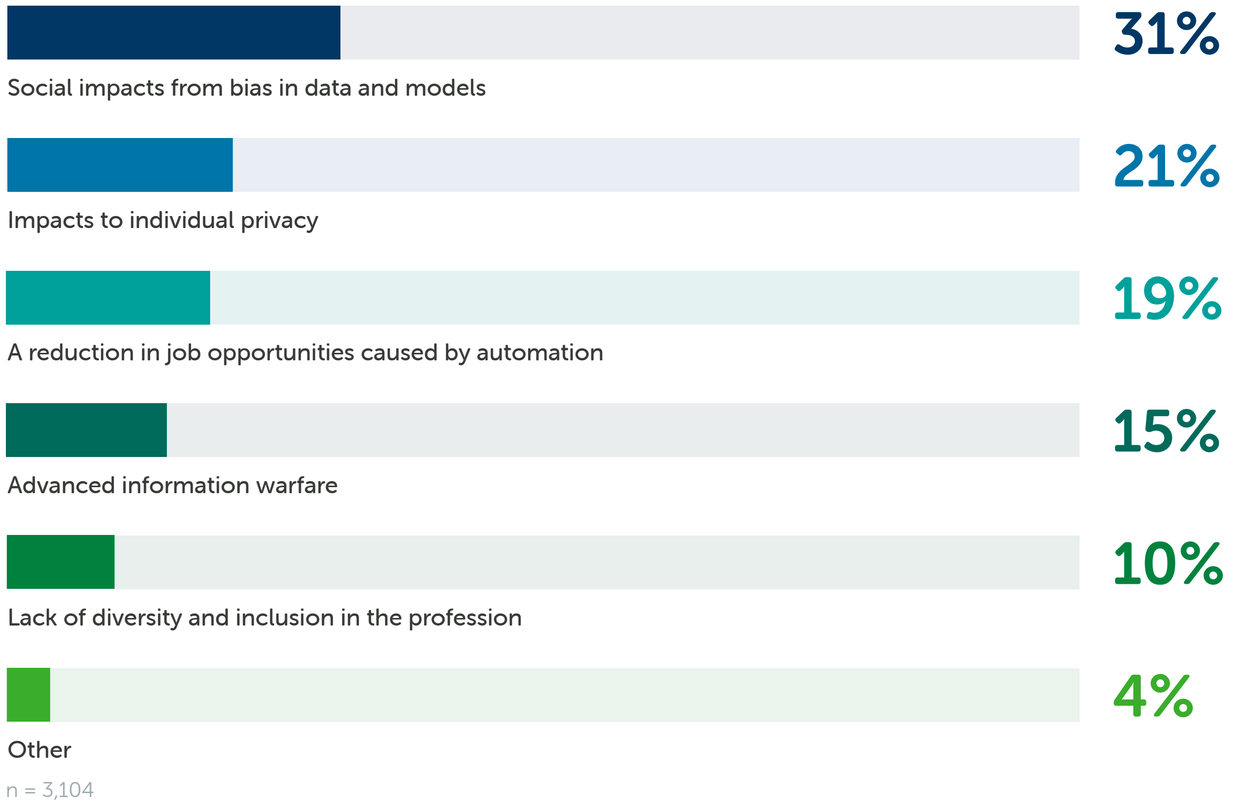
said finding an internship is their biggest obstacle to obtaining experience, followed by 20% who said learning a new language.

The data field is competitive and saturated with a variety of titles. Therefore, students need to learn about the technical and soft skills required to succeed and the differences between job functions in data-related fields to better chart a potential course for a career. In addition, these functions and expectations may change between mature and smaller organizations. Understanding the pain points of these different roles, their overall job satisfaction, and their involvement in business decisions can guide students toward any additional courses or online learning to improve their abilities.

BIG QUESTIONS

As we think about the future, we want to know the sentiments from our respondents toward trends, potential opportunities, and any setbacks to growth. Given the influential role ML and AI play in business and society, data professionals grapple with big questions that often impact how the field is evolving.

What our respondents felt was the biggest problem to tackle in AI/ML



31% of respondents believe the most significant problem to tackle in AI/ML today is the social impacts caused by bias in data and models. However, not all organizations and educational institutions are taking the necessary steps to tackle this issue.

Our respondents said the social impacts of bias in data and models are the most significant problem in AI/ML. Therefore, we asked about their teams' steps to mitigate bias and ensure model explainability.

Is your data science team planning to take any steps to ensure fairness and mitigate bias or to address model explainability?

Fairness and bias mitigation

30% No, and we are not planning to

30% Yes, we are planning to in the next 12 months

10% Yes, we have already implemented at least one step

30% I don't know

n = 2,135

Model explainability and interpretability

31% No, and we are not planning to

31% Yes, we are planning to in the next 12 months

10% Yes, we have already implemented at least one step

27% I don't know

n = 2,135

While only 10% of respondents indicated their organizations have already implemented a solution to ensure fairness and mitigate bias, it's encouraging to see 30% said they plan to implement steps in 12 months, a 7% increase from 2020 (23%).

Similarly, while 31% of respondents stated their organizations also don't have any plans to ensure model explainability and interpretability, 41% said they plan to take steps in the next 12 months or have already implemented at least one step.

While steps to mitigate bias and ensure fairness are also present in higher education, only 17% of educators responded that they are teaching students about ethics in data science and ML, and only 16% of students stated they are being taught ethics in data science and ML in preparation to enter the field. Similarly, only 22% of educators and 24% of students said that bias in AI/ML/data science is taught frequently in class or lectures – 41% of educators and 45% of students responded that it is rarely or never taught. To move the needle forward for supporting fairness in data, educational institutions and educators must also make this a priority.

When breaking down the survey results further by organization size, organizations of 10,001+ employees had the most significant percentage of respondents say they had already implemented at least one step to help ensure fairness and mitigate bias in data sets and models. Since large organizations often require additional resources to be at the leading edge of industry work, they can allocate resources to problems like combating bias in AI. However, when employees of large organizations were asked about their plans to mitigate bias, the most common response was still, "I don't know." This suggests that there's room to educate all employees about combating bias in AI/ML. In addition, more interdepartmental work on these topics will allow employees to participate in discussions about their companies' policies in these areas, advocating for further work where needed.

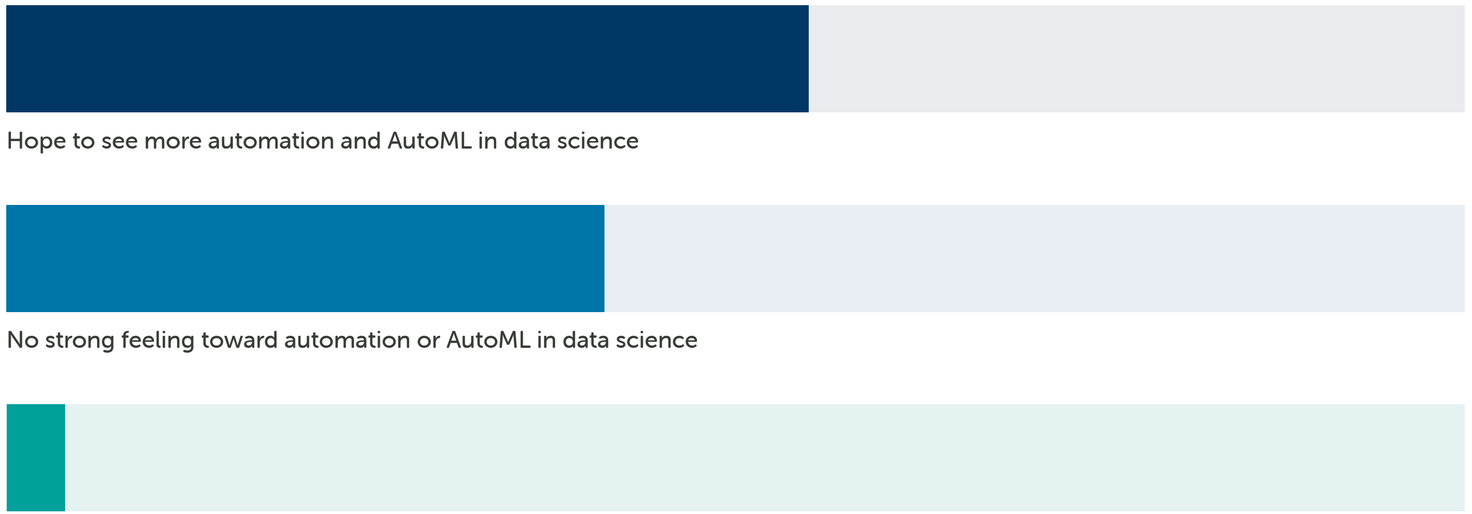
Ensuring Fairness, Mitigating Bias, and Model Explainability

Data is always going to have some inherent bias due to natural limitations and constraints. Because all data has an unintentional bias, data should be examined from this lens, even if you think it is unbiased and fair. Considering social impacts from bias in data and models is one of the biggest concerns for organizations, it's essential to ask questions like, "How did we choose which data to collect?" and, "Which data points may be excluded as a result?" It is also beneficial to open up access to data for review by multiple parties who bring different perspectives to the table.

With that review process, it's also important to consider explainability and interpretability. While the number of organizations implementing a solution slipped from last year, this could be because of the pandemic. Of those who saw a decrease in resources due to COVID-19, 35% of responses indicated that project timelines were on hold indefinitely or for an extended period. We hope to see that as our world changes, mitigating bias, ensuring fairness, and model explainability continues to progress.

Positive sentiment toward AutoML in data science

What is your sentiment toward automation or AutoML, the process of automating tasks involved in applying machine learning to real-world problems, in data science?

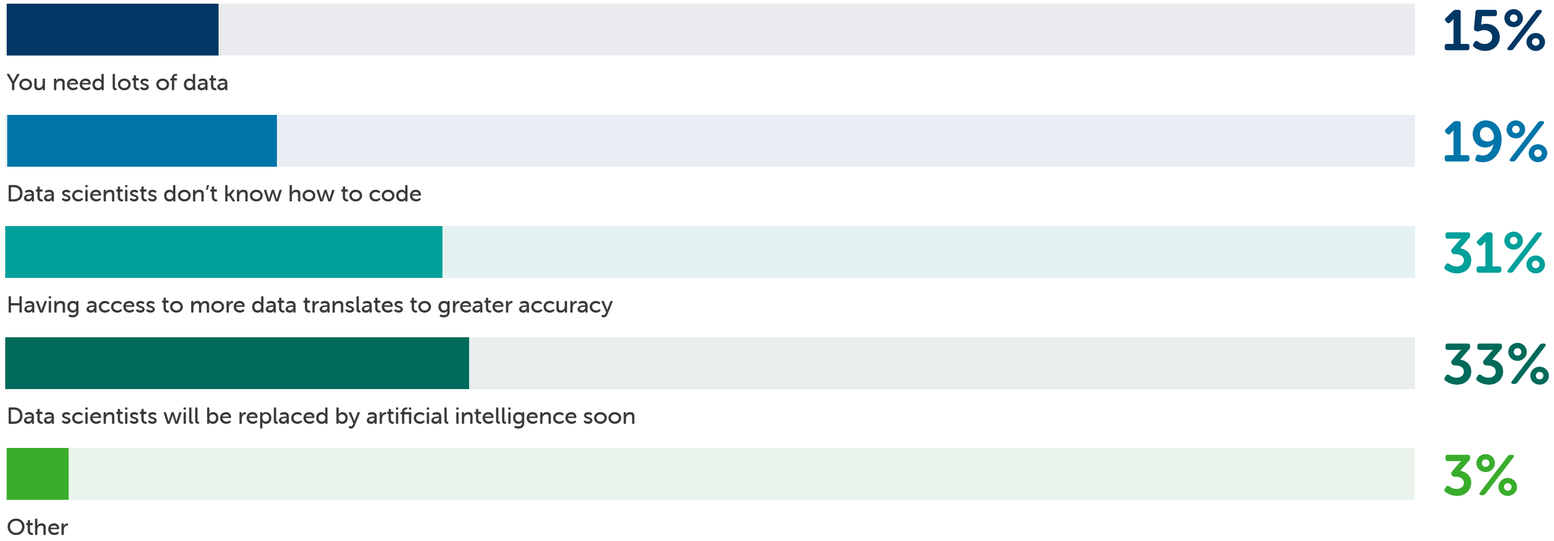


n = 3,104

55% of respondents hope to see more automation and AutoML in data science, while only 4% are concerned with how automation will impact data science.

A common theme in the news today is that automation is taking over and will eventually replace human workers. However, results show that automation is welcomed in the data science sector and isn't viewed as a competitor but rather a complementary tool to practitioners.

What is the biggest myth about data science?



n = 3,104

Additionally, when asked about data science myths, 33% of respondents replied, saying the biggest myth is that data scientists will be replaced by AI soon. On the contrary, there is an opportunity for AI and automation to help with easily repeatable tasks, to free up more resources for work that requires human intervention, interpretation, and problem solving. Automation will allow individuals to develop more complex models or algorithms and spend less time on routine work that doesn't need to rely on personalization or a high level of detail.

A majority of data science professionals spend their days prepping and cleansing data. While automating some data prep processes, such as data anonymization, will help free up time, it's important to remember that data preparation helps data scientists better understand the data and its limitations in order to build better models.

Similarly, there is a myth in data science that says quantity over quality is best. 31% of respondents said they believe the biggest myth in data science is that having access to more data translates to greater accuracy. The old adage is still valid at the end of the day, "garbage in, garbage out." Humans need to be in the mix to provide the context to understand what is considered "good" and bring their different perspectives to the table.

33%

of respondents believe the biggest myth is data scientists will be replaced by AI soon.

LOOKING AHEAD

As we prepare for what's next in the field, we've outlined four themes for enterprises to focus on.

1 PYTHON WILL CONTINUE TO DOMINATE THE DATA SCIENCE FIELD AND BEYOND.

Python is a top programming language and is at the core of data science and scientific computing. It makes the entry to data science simple for students and beginners and is used by experts around the world. In addition, its open-source nature allows for continuous innovation in the supporting ecosystem of libraries and packages, making the experience a positive one for makers and users alike.

Additionally, the Python community is very active. There are countless developer forums for questions and troubleshooting and many libraries to enable diverse workstreams. And with ongoing development in the open-source ecosystem, there are countless ways for people to get involved.

Python already has a large and diverse user base that we see becoming an organizational standard. Our CEO and co-founder, Peter Wang, believes Python will be the successor to Excel since it lowers the barrier to entry in data science across multiple business functions. By bridging the gap between technical roles and business functions, we can improve data literacy and the way businesses use data.

2 ENTERPRISES ARE READY TO CONTRIBUTE TO OPEN-SOURCE INNOVATION.

Open-source software adoption is high. 87% of survey respondents in commercial organizations indicated they are already using open-source technology today. It is no secret that open-source maintainers and contributors have traditionally been hobbyists, but now, we're seeing more employers encouraging their teams to contribute to open-source innovation. Through resources like increased funding related to open-source project development, additional time dedicated to contributing, and team members dedicated to contributing to open-source projects, we can expect even more innovation through open-source software. Even as the past year has presented its share of challenges, only 18% of survey respondents said that employer support for open source decreased due to COVID-19 or other factors.

However, with the rapid growth of open source adoption comes challenges. Open-source software presents some potential risks and vulnerabilities that make some companies anxious and drive decisions to use vendor software instead. Fear of CVEs, possible exposures, or risks are the main obstacles to adoption. But of those who have adopted, a quarter of respondents said their organization does not take any steps to secure its open-source data science and machine learning pipeline, despite the availability of such solutions.

As recent news cycles and the recent US Executive Order have shown, cybersecurity attacks pose a serious threat to organizations today. For companies that use open source or third-party software, supply chain verification of their code will become increasingly important. It is possible to integrate open-source software into a business's tech stack in a secure way by following steps such as regular CVE database checks, timely patching of vulnerabilities, and working with securely-built packages from the outset.

3 SENTIMENT TOWARD AUTOMATION WILL CONTINUE TO GROW.

Despite themes in the media suggesting automation will 'take over,' it's actually welcomed by data practitioners and isn't seen as a competitor. AutoML is valued for its ability to quickly and efficiently tune many hyperparameters, help choose the best model types to solve specific problems and enable non-experts to train ML models. Automation is seen as a complement to work. Data scientists aren't worried because they recognize how many aspects of their job still require expert human judgment that technology can't replicate. Additionally, since most AutoML frameworks today are fairly generic, and as organizations look to integrate data science more closely with their particular business domains, they'll need data scientists to carry out more complex and specific work. AutoML can help free up data scientists' time spent on some of their simpler tasks so that they can concentrate on these higher-level challenges.

4 PREVENTING BIAS AND DEVELOPING ETHICAL DATA SCIENCE IS CRITICAL.

For all the benefits that AI/ML have brought to business and society, we also must consider the problems and challenges these tools have introduced and the steps we can take to mitigate those issues. When asked what they view as the most significant problem to tackle in the AI/ML area today, the most popular answer was the social impacts caused by bias in data and models. It's encouraging to see an increase in organizations planning to implement at least one step in the next year. However, most individuals still don't know if their organization is taking any steps to ensure fairness and mitigate bias. Additionally, students and universities aren't prioritizing ethics and mitigating bias, so it's often an afterthought in the field.

Although bias in AI/ML data and models is top of mind for many practitioners, the gap between awareness of the issue and active implementation of solutions suggests that there's still work to be done in this area.

As we consider what's next for the data science field, we're building a dedicated home for our community to gain expert insights, connect with experts, and more. Check out Anaconda Nucleus to continue your learning journey.

ABOUT ANACONDA

With more than 25 million users, Anaconda is the world's most popular data science platform and the foundation of modern machine learning. We pioneered the use of Python for data science, champion its vibrant community, and continue to steward open-source projects that make tomorrow's innovations possible. Our enterprise-grade solutions enable corporate, research, and academic institutions around the world to harness the power of open-source for competitive advantage, groundbreaking research, and a better world.

Visit https://www.anaconda.com to learn more.

ANACONDA.